机器学习
第2节

涉及知识点：
线性回归、梯度更新方式、线性回归矩阵形式、泛线性模型、最大似然估计、逻辑斯谛回归、分类指标、逻辑斯谛回归的实践

线性模型

张伟楠 – 上海交通大学

线性回归

张伟楠 – 上海交通大学

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型 (Conditional Models)
- 分类
  - 确定性判别模型： $y = f_\theta(x)$
  - 概率判别模型： $p_\theta(y|x)$

本节集中介绍线性判别模型 (linear regression)

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型(Conditional Models)
- 分类
  - 确定性判别模型：$y = f_\theta(x)$
  - 概率判别模型：$p_\theta(y|x)$

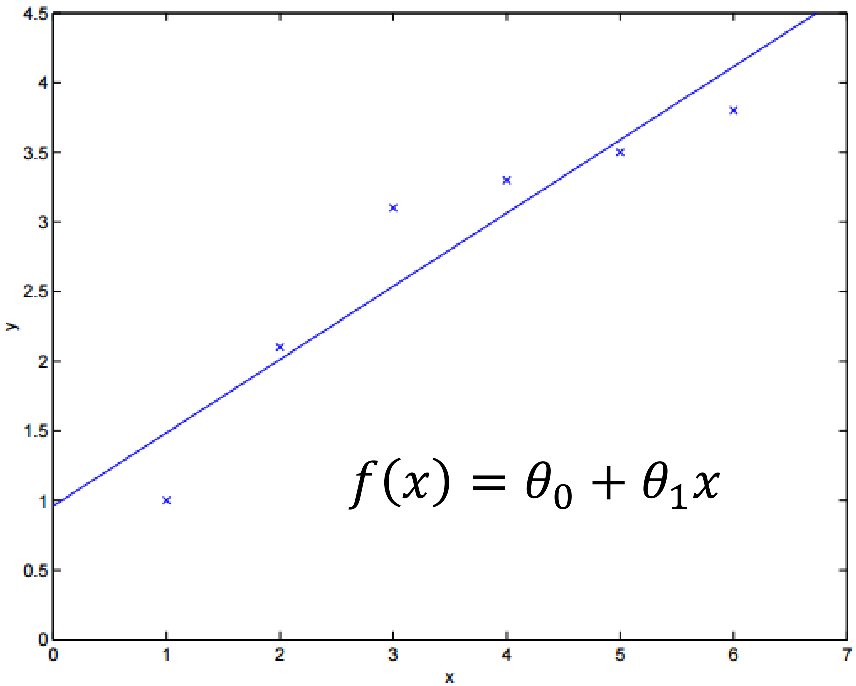
线性判别模型 (linear regression)

$$y = f_\theta(x) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j = \theta^\top x$$

$$x = (1, x_1, x_2, \cdots, x_d)$$

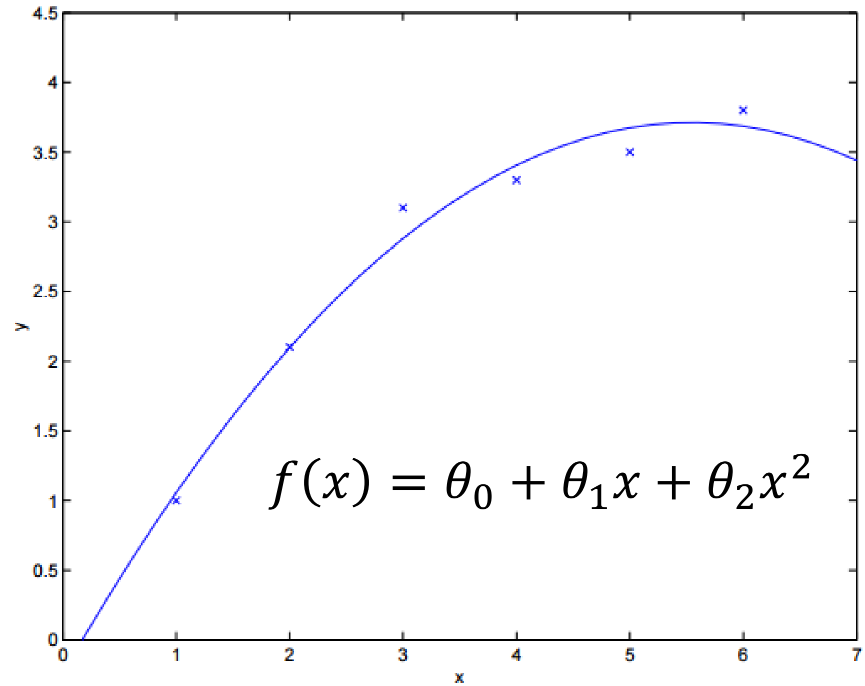
线性回归

- 一维的线性回归和二次回归（都是线性模型）



$$f(x) = \theta_0 + \theta_1 x$$

线性回归



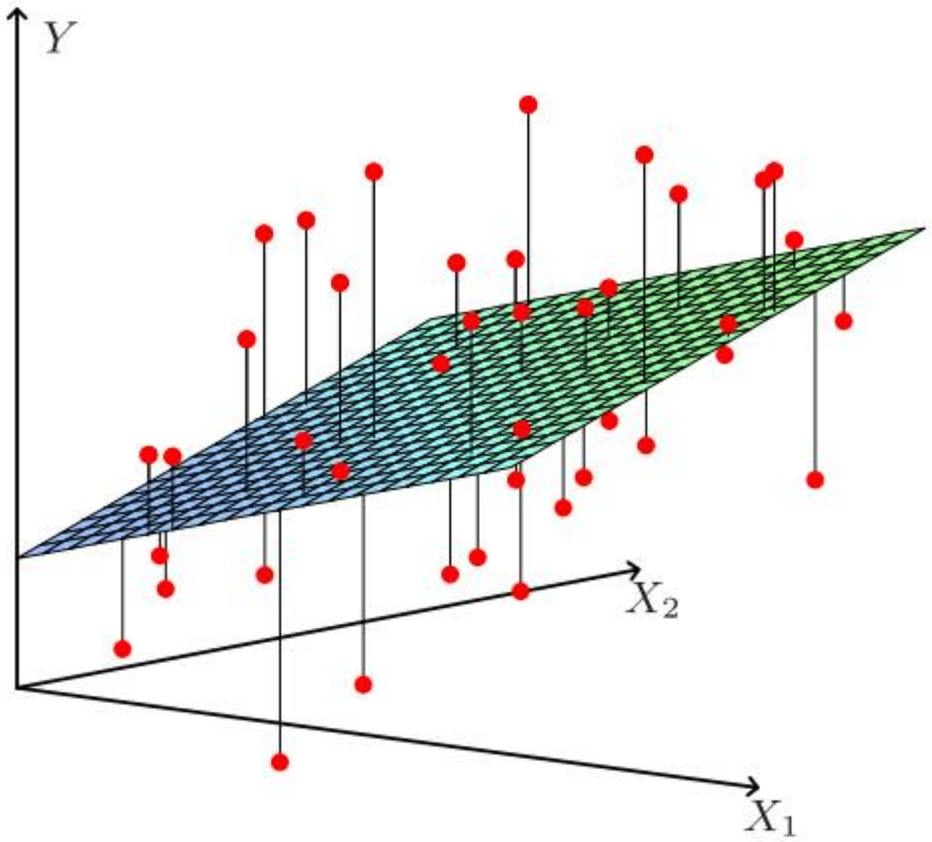
$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

二次回归
(一种广义线性模型)

线性回归

- 二维的线性回归模型

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



学习目标

- 使预测值和真实值的距离越近越好

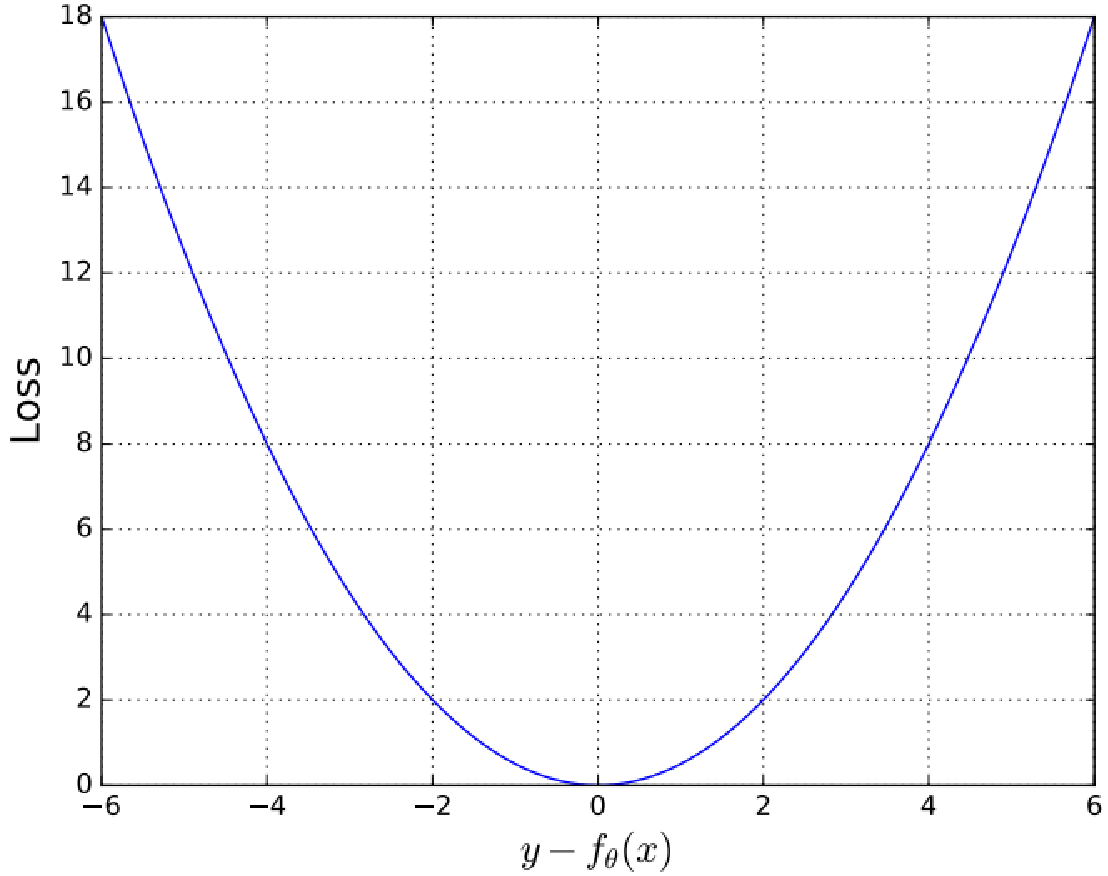
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f_\theta(x_i))$$

- 损失函数$\mathcal{L}(y_i, f_\theta(x_i))$ 测量预测值和真实值之间的误差，越小越好
- 具体损失函数的定义依赖于具体的数据和任务
- 最广泛使用的损失回归函数：平方误差(squared loss)

$$\mathcal{L}(y_i, f_\theta(x_i)) = \frac{1}{2}\left(y_i - f_\theta(x_i)\right)^2$$

平方误差

$$\mathcal{L}\big(y_i, f_\theta(x_i)\big) = \frac{1}{2}\big(y_i - f_\theta(x_i)\big)^2$$

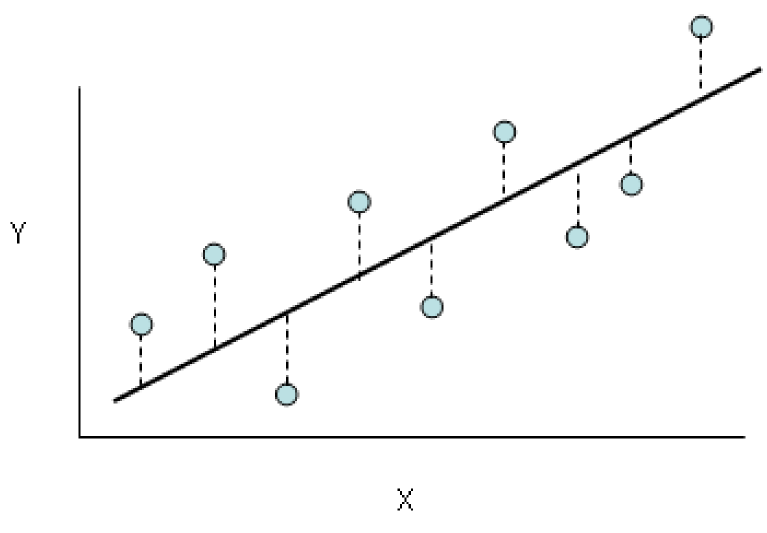


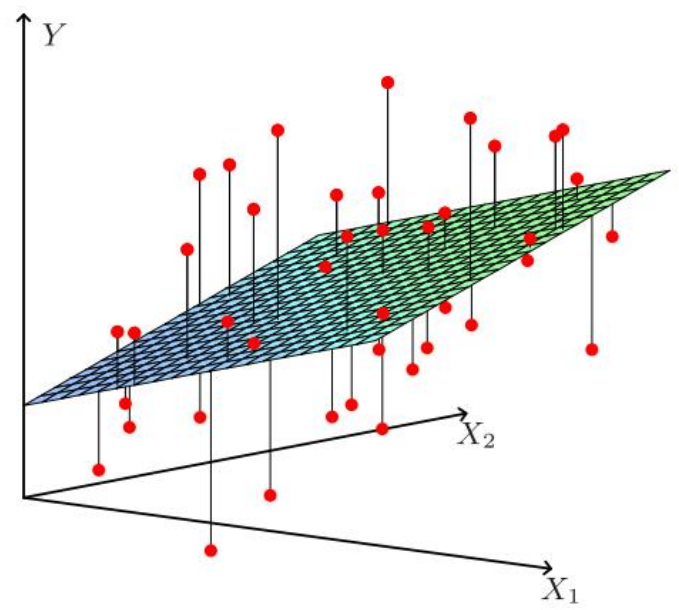
- 对预测误差大的有更大的惩罚
- 容忍很小的预测误差
  - 观测误差等
  - 提升模型的泛化能力

最小均方误差回归

- 优化目标是最小化训练数据上的均方误差

$$J_\theta = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)^2 \qquad \min_\theta J_\theta$$

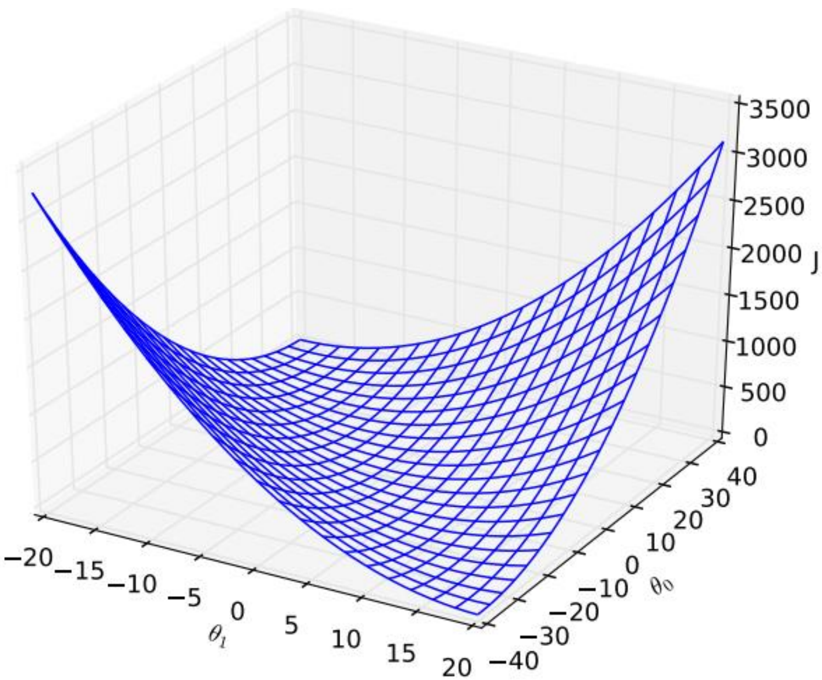


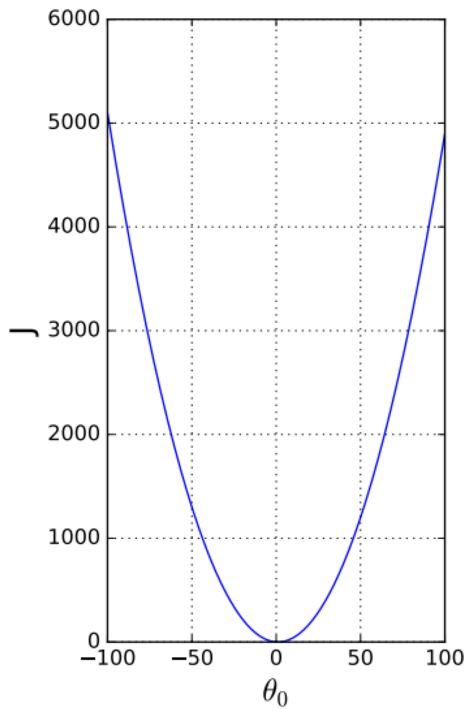


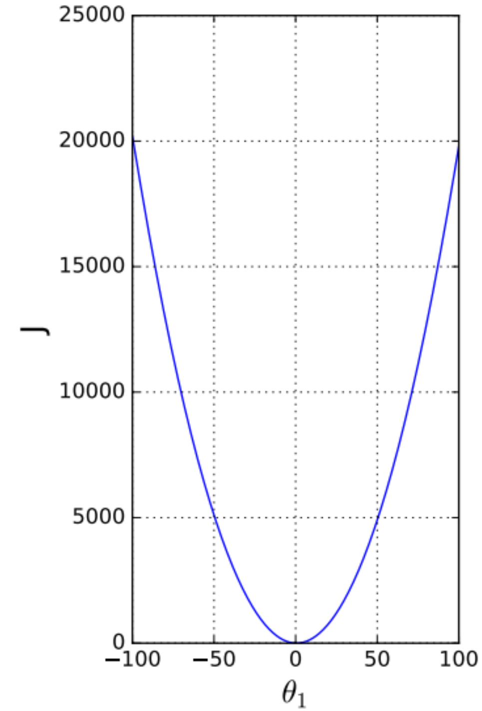
最小化目标函数

- 举一个$N=1$的简单示例，对于数据点$(x,y)=(2,1)$

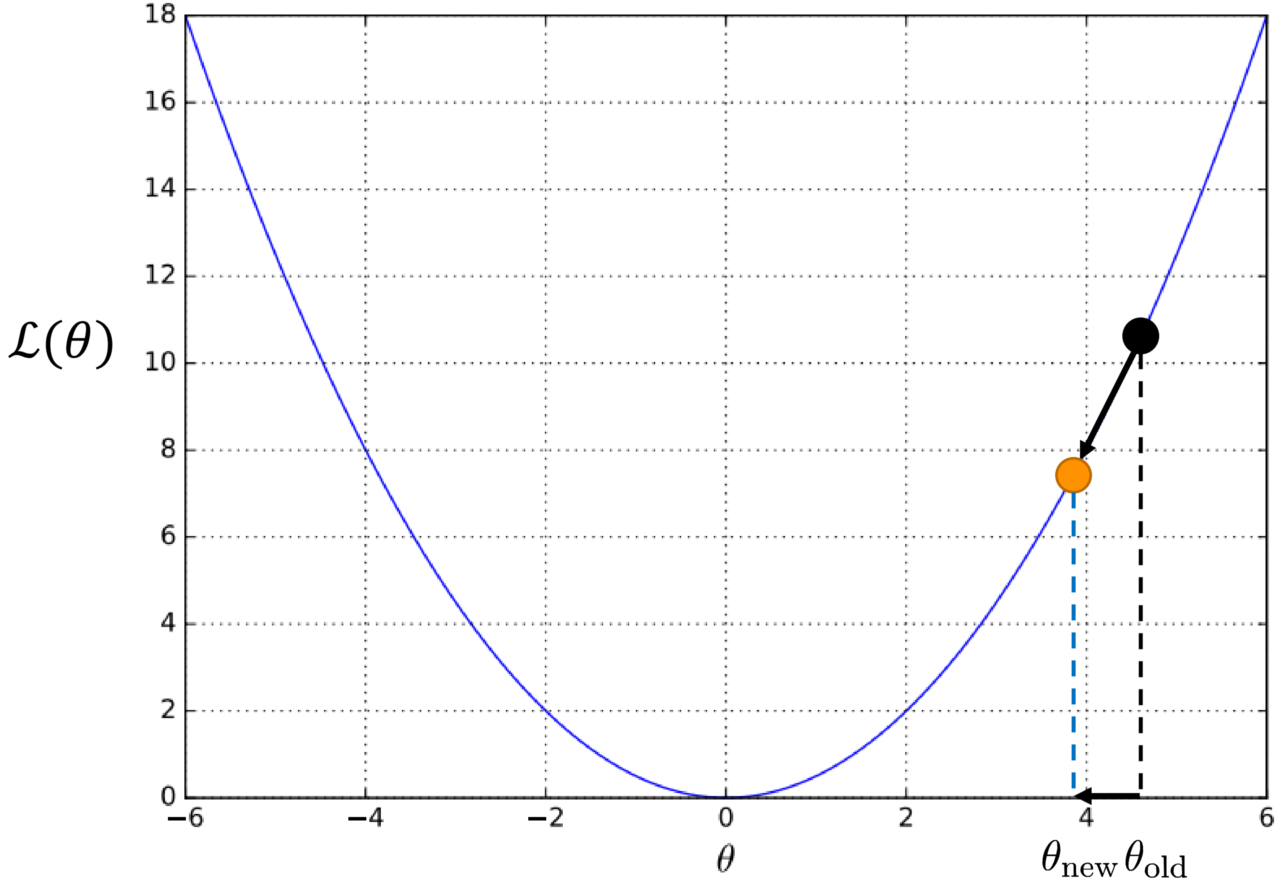
$$J(\theta)=\frac{1}{2}(y-\theta_0-\theta_1 x)^2=\frac{1}{2}(1-\theta_0-2\theta_1)^2$$







梯度学习方法



$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

梯度更新方式

张伟楠 – 上海交通大学

目录

Contents

01

批量梯度下降

批量梯度下降

- 优化目标

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)^2 \qquad\qquad \min_\theta J(\theta)$$

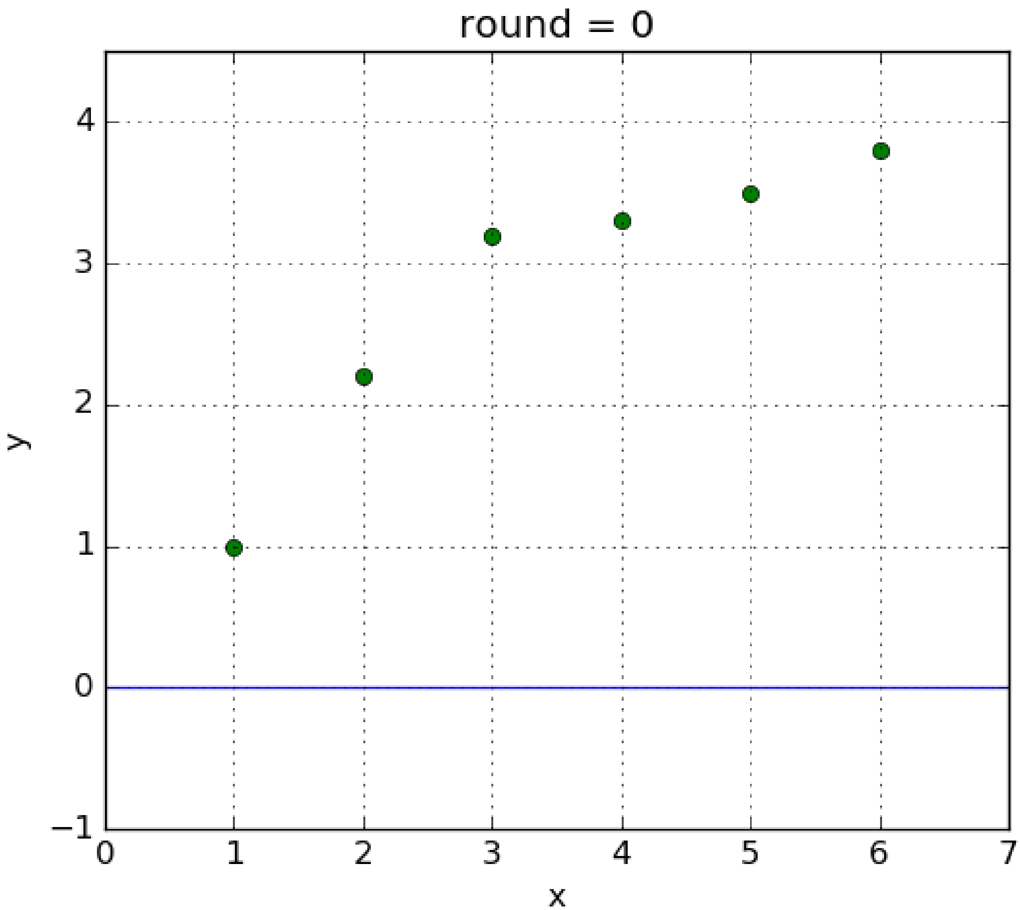
- 根据整个批量数据的梯度更新参数 $\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta\frac{\partial J(\theta)}{\partial\theta}$

$$\frac{\partial J(\theta)}{\partial\theta} = -\frac{1}{N}\sum_{i=1}^{N}\left((y_i - f_\theta(x_i))\frac{\partial f_\theta(x_i)}{\partial\theta}\right)$$

$$= -\frac{1}{N}\sum_{i=1}^{N}(y_i - f_\theta(x_i))\,x_i$$

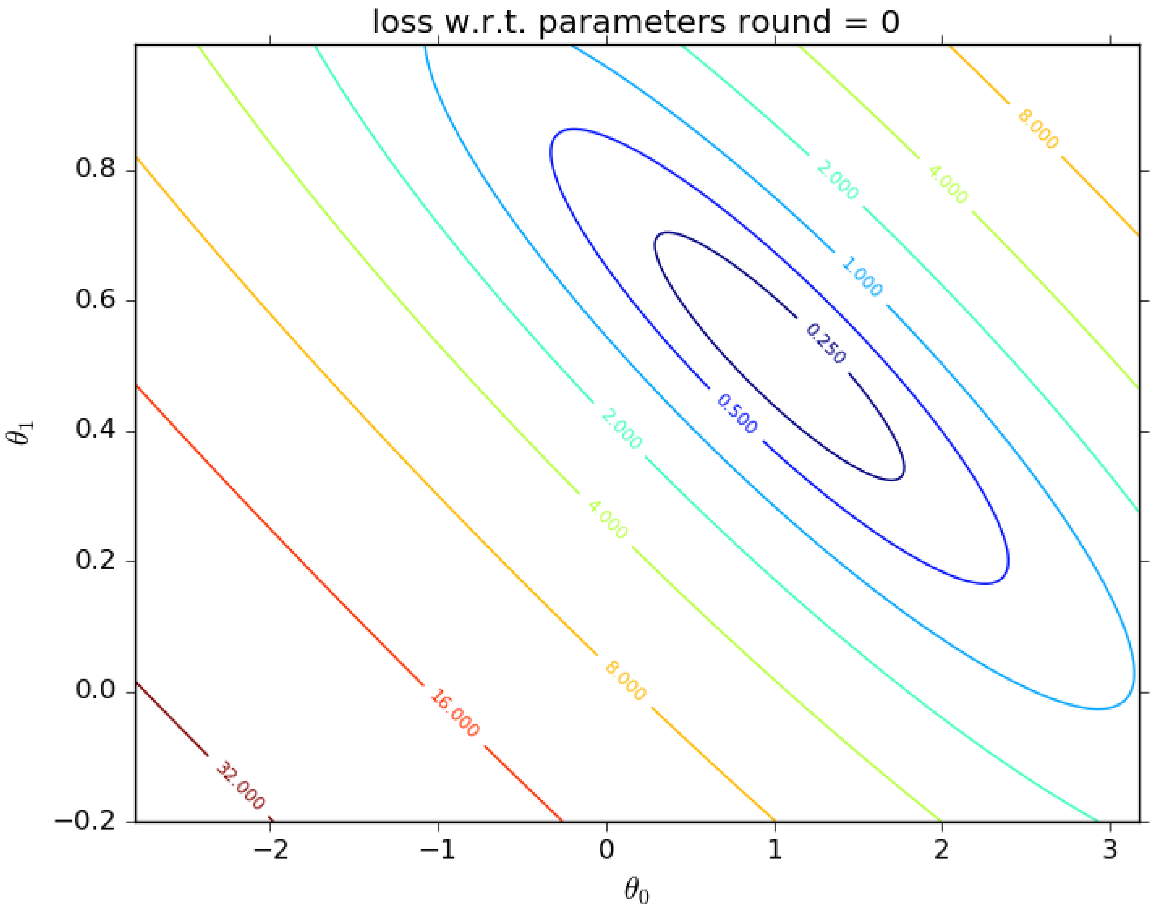
$$\theta_{\text{new}} = \theta_{\text{old}} + \eta\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)x_i$$

线性模型学习展示 - 函数曲线



$$f(x) = \theta_0 + \theta_1 x$$

线性模型学习展示 – 参数改变



批量梯度更新

02

随机梯度下降

随机梯度下降

- 优化目标

$$J^{(i)}(\theta) = \frac{1}{2}\left(y_i - f_\theta(x_i)\right)^2 \qquad \min_\theta \frac{1}{N}\sum_i J^{(i)}(\theta)$$

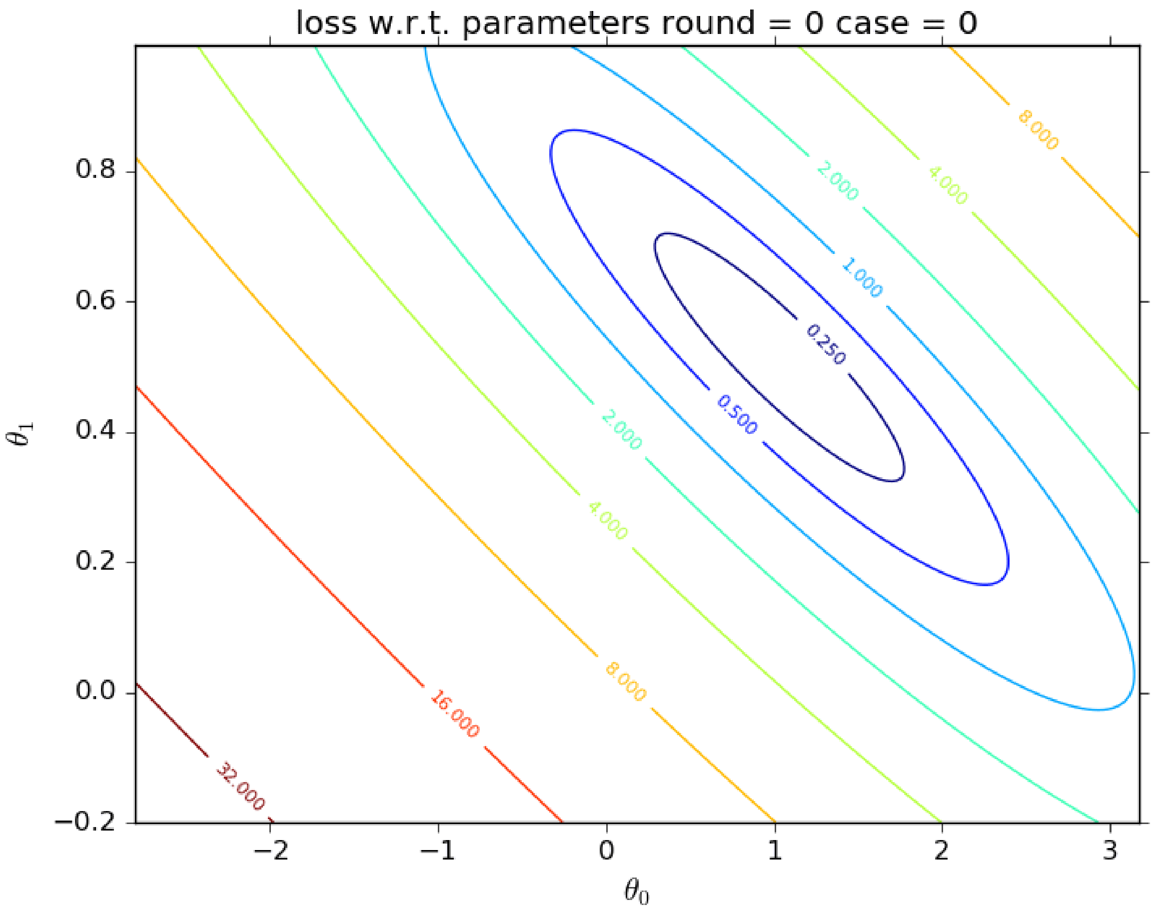
- 根据整个批量数据的梯度更新参数 $\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J^{(i)}(\theta)}{\partial \theta}$

$$\frac{\partial J^{(i)}(\theta)}{\partial \theta} = -\left(y_i - f_\theta(x_i)\right)\frac{\partial f_\theta(x_i)}{\partial \theta}$$
$$= -(y_i - f_\theta(x_i))\, x_i$$

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta\left(y_i - f_\theta(x_i)\right)x_i$$

- 对比批量梯度下降
  - 更快地更新参数(优点)
  - 学习中不确定性或震荡(缺点)

线性模型学习展示 – 参数改变



随机梯度更新

03

小批量梯度下降

小批量梯度下降

算法思想

批量梯度下降和随机梯度下降的结合

训练步骤

- 将整个训练集分成 K 个小批量（mini-batches）

$$\{1,2,3,\cdots,K\}$$

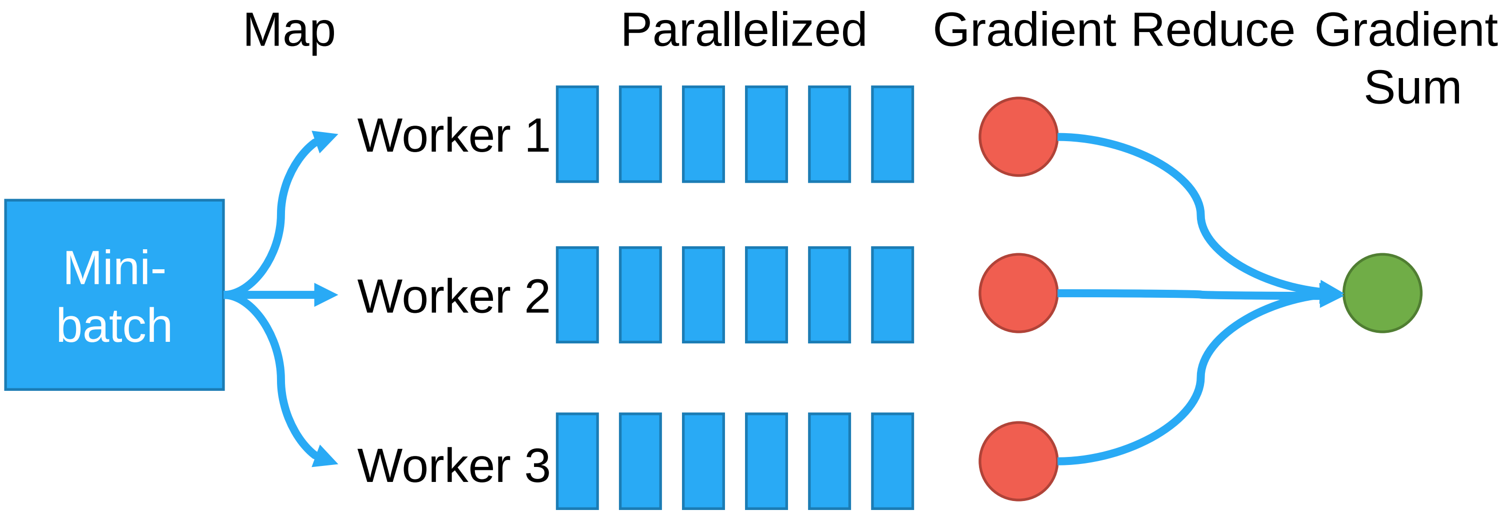
- 对于每一个小批量 k, 做一步批量下降来降低

$$J^{(k)}(\theta)=\frac{1}{2N_k}\sum_{i=1}^{N_k}\left(y_i-f_\theta(x_i)\right)^2$$

- 对于每一个小批量，更新参数 $\theta_{\text{new}} = \theta_{\text{old}}-\eta\frac{\partial J^{(k)}(\theta)}{\partial\theta}$

小批量梯度下降

- 结合了批量梯度下降和随机梯度下降的优点
  - 批量梯度下降的优秀的稳定性
  - 随机梯度下降的快速更新
- 小批量梯度下降很适合使用在并行化计算中
  - 将每个小批量数据进一步切分，每个线程分别计算梯度，最后再加和这些梯度

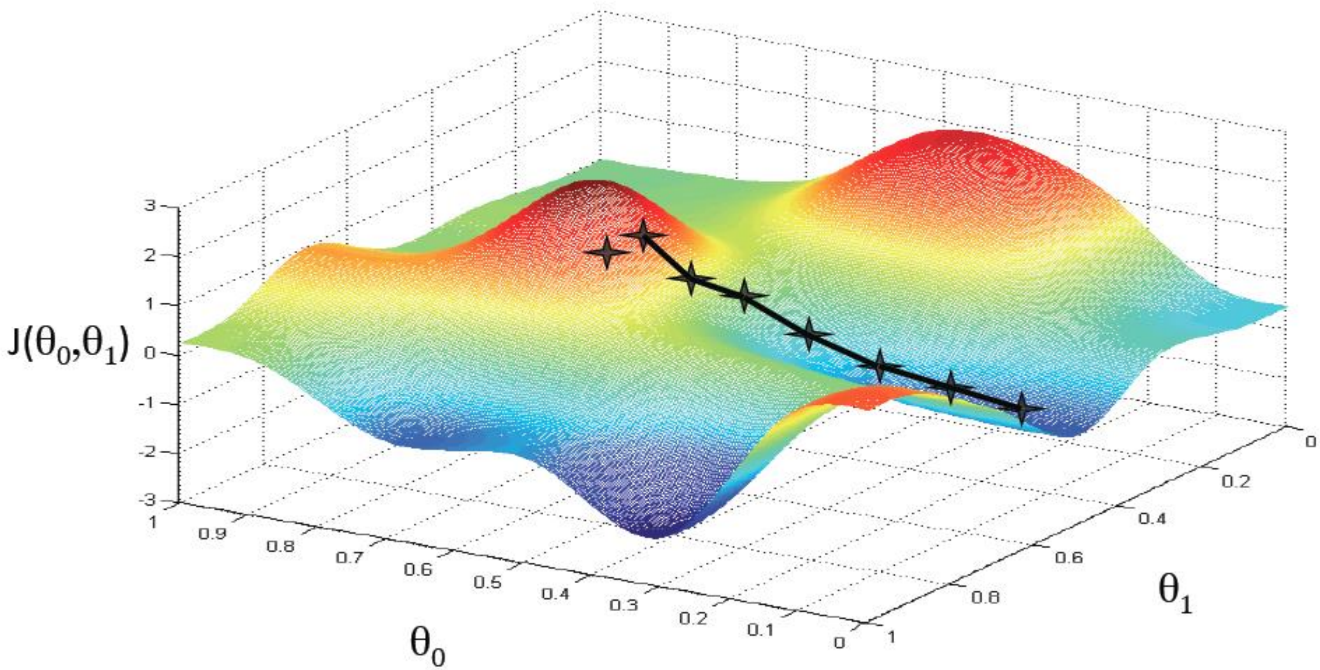


04

基本搜索步骤

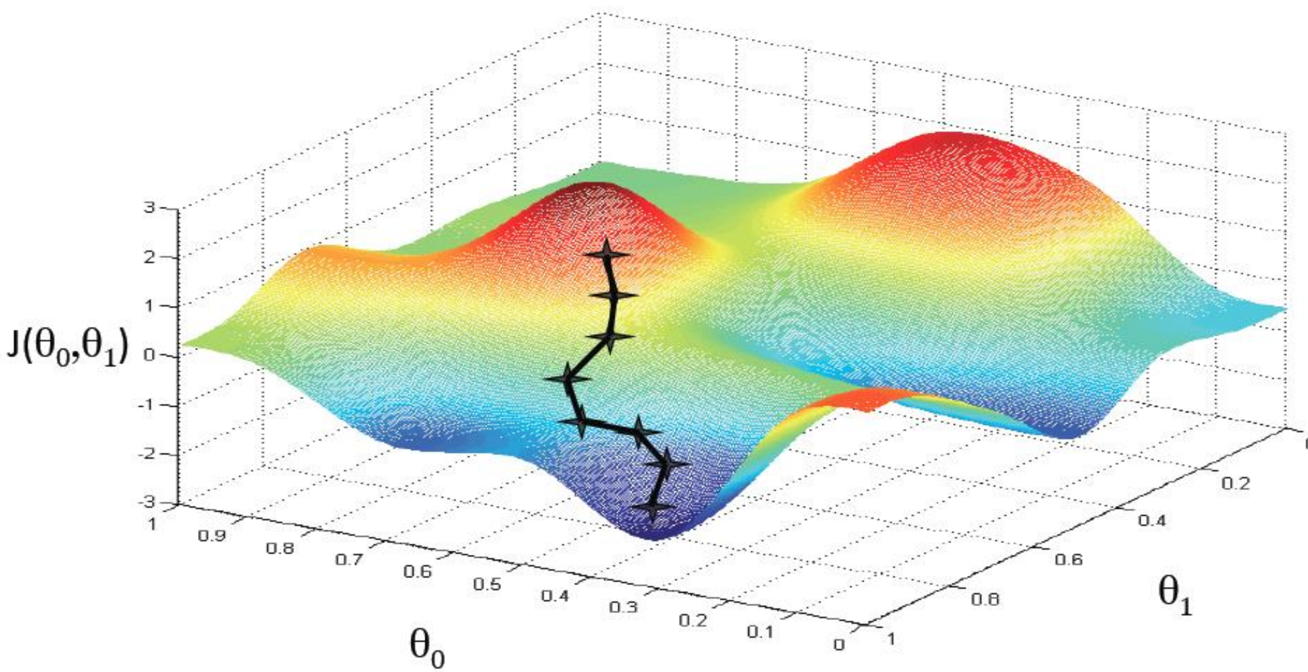
基本搜索步骤

- 随机选择一个参数初始化 θ
- 根据数据和梯度算法来更新 θ
- 直到走到局部一个最小区域(local minimum)

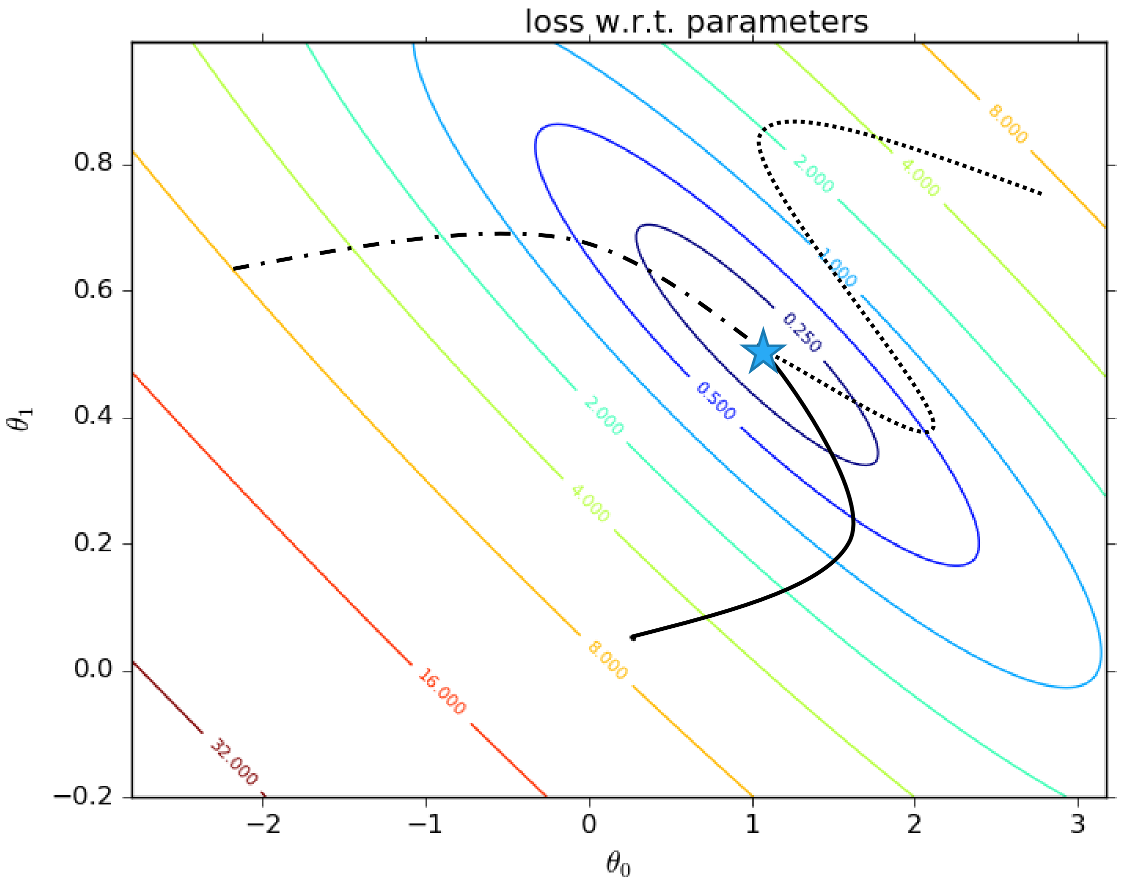


基本搜索步骤

- 随机选择一个新的参数初始化 θ
- 根据数据和梯度算法来更新 θ
- 直到走到局部一个最小区域(local minimum)



凸优化目标函数具有唯一最小点



- 不同的初始化参数最终也会学习到相同的最优值

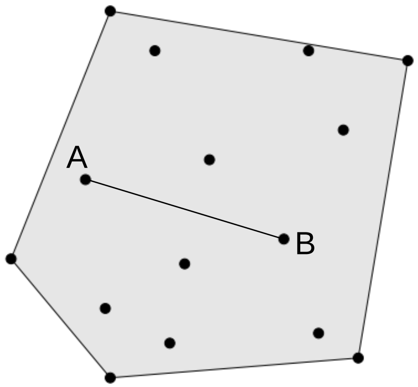
凸集 (Convex Set)

凸集的定义

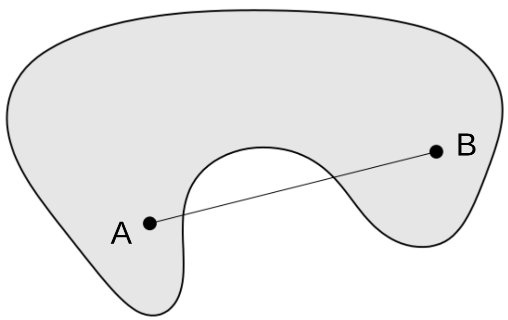
一个点集S被称为凸集，当且仅当该S里的任意两点A和B的连线上任意一点同样属于S

$$tx_1 + (1-t)x_2 \in S$$

$$\text{for all } x_1, x_2 \in S, 0 \le t \le 1$$



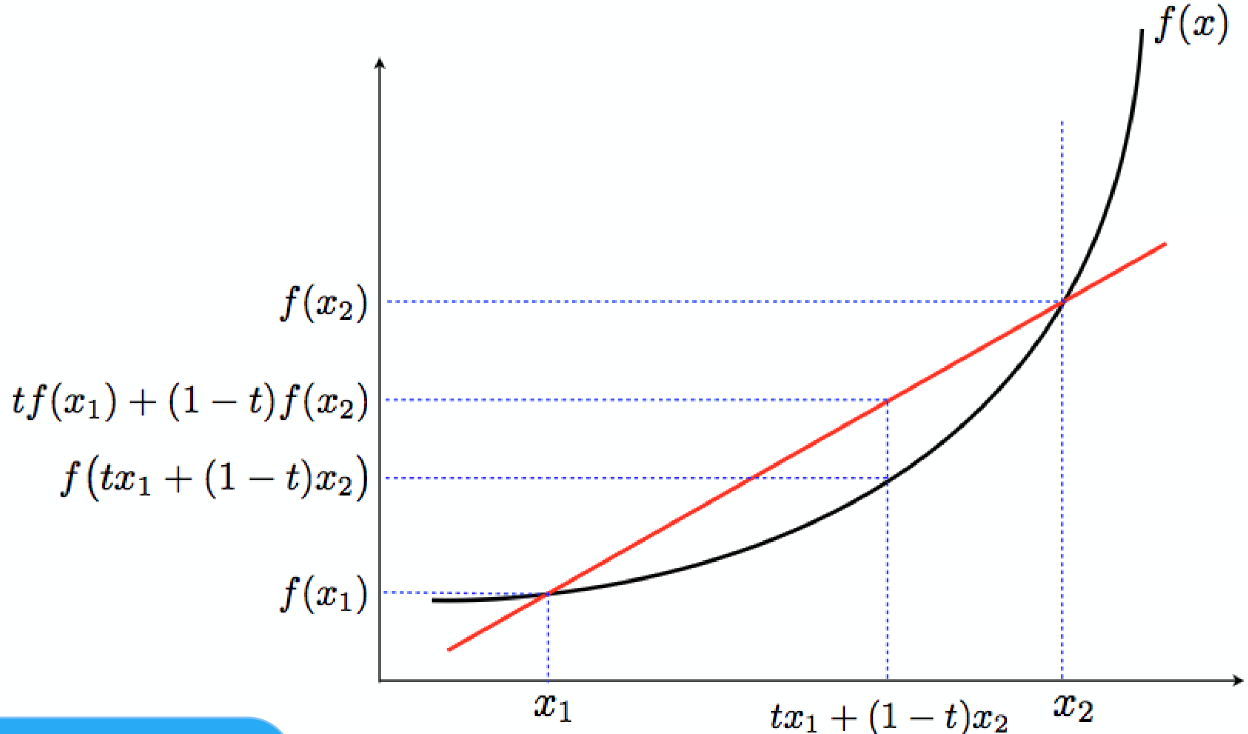
凸集



非凸集

Boyd, Stephen, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

凸函数 (Convex Function)



凸函数的定义

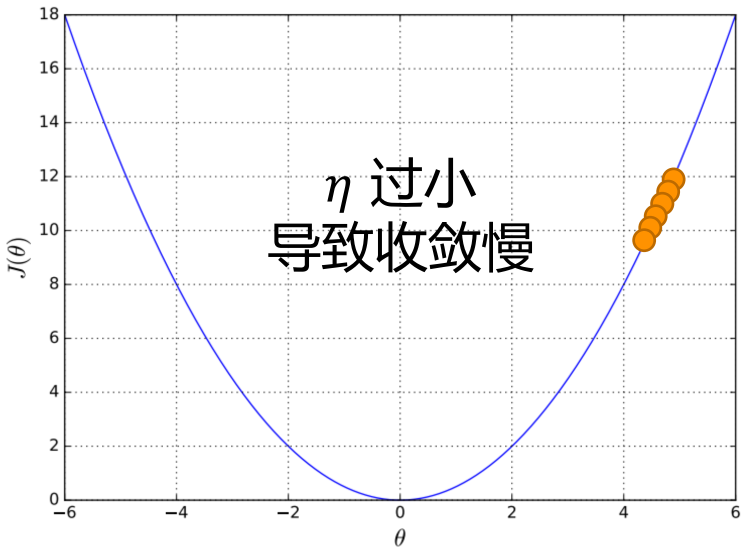
$f: \mathbb{R}^n \to \mathbb{R}$ 是凸函数: **dom** f 是一个凸集，并且满足

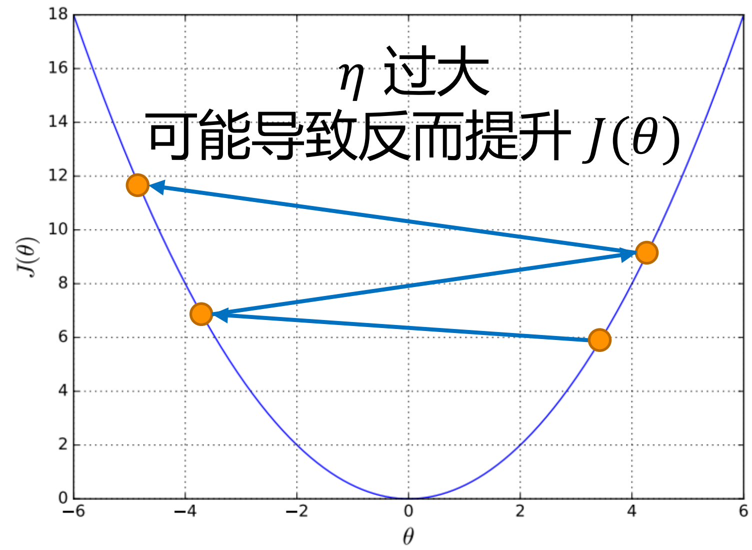
$$f(tx_1+(1-t)x_2) \le tf(x_1)+(1-t)f(x_2)$$

$$\forall x_1, x_2 \in \mathbf{dom} f, 0 \le t \le 1$$

学习率的选择

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J(\theta)}{\partial \theta}$$





- 初始点可能距离最优点太远，从而导致收敛速度慢

- 可能越过最优点
- 可能无法收敛
- 甚至可能发散

- 要检查梯度下降是否有效工作，可以打印出每几个迭代得到的损失 $J(\theta)$, 如果发现 $J(\theta)$ 并没有正常地下降，调整学习率 η

线性回归矩阵形式

张伟楠 – 上海交通大学

从代数视角来看线性回归

训练数据矩阵

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(n)} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & & x_d^{(2)} \\ \vdots & & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} \quad \text{参数 } \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \text{ 标签 } \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

预测

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{x}^{(1)}\boldsymbol{\theta} \\ \boldsymbol{x}^{(2)}\boldsymbol{\theta} \\ \vdots \\ \boldsymbol{x}^{(n)}\boldsymbol{\theta} \end{pmatrix}$$

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

线性回归的矩阵形式

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

最优参数求解

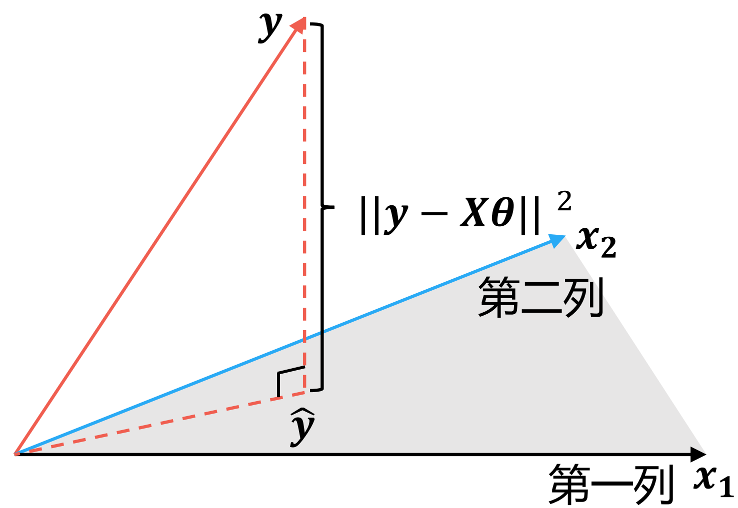
$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 &\rightarrow \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = 0 \\ &\rightarrow \boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} \\ &\rightarrow \widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\end{aligned}$$

线性回归的矩阵形式

预测值

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

$\boldsymbol{H}$:帽子矩阵



几何解释

- 数据矩阵的列向量$[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_d]$张成一个 $\mathbb{R}^n$ 上的子空间
- $\boldsymbol{H}$就是将标签向量$\boldsymbol{y}$投影到该子空间的映射

$$\boldsymbol{X} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & & x_d^{(1)} \\ & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_d] \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$X^T X$为奇异矩阵的情况

- 当数据矩阵的一些列向量线性相关时
  - 例如$\boldsymbol{x}_2 = 3\boldsymbol{x}_1$
- $\boldsymbol{X}^\top \boldsymbol{X}$为奇异矩阵，所以$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$无法被直接计算。

解决方案

- 正则化（Regularization）
- $J(\boldsymbol{\mu}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\theta})^\top (\boldsymbol{y} - \boldsymbol{X\theta}) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$

带正则项的线性回归矩阵形式

优化目标

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}$$

最优参数求解

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \to -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta} = 0$$

$$\to \boldsymbol{X}^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\theta}$$

$$\to \widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

泛线性模型

张伟楠 – 上海交通大学

回顾：线性回归

训练数据矩阵

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} \quad 参数\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \quad 标签\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

预测

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{x}^{(1)}\boldsymbol{\theta} \\ \boldsymbol{x}^{(2)}\boldsymbol{\theta} \\ \vdots \\ \boldsymbol{x}^{(n)}\boldsymbol{\theta} \end{bmatrix}$$

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

回顾：线性回归的矩阵形式

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

最优参数求解

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \ \rightarrow \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = 0$$

$$\rightarrow \boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}$$

$$\rightarrow \widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

泛线性模型

相关关系(Dependence)

$$y = f(\theta^\top \phi(x))$$

- 特征映射函数 $\phi(x)\colon \mathbb{R}^d \mapsto \mathbb{R}^h$
- 特征映射矩阵 $\Phi_{n\times h}$

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(x^{(1)}) \\ \phi(x^{(2)}) \\ \vdots \\ \phi(x^{(i)}) \\ \vdots \\ \phi(x^{(n)}) \end{bmatrix} = \begin{bmatrix} \phi_1(x^{(1)}) & \phi_2(x^{(1)}) & \cdots & \phi_h(x^{(1)}) \\ \phi_1(x^{(2)}) & \phi_2(x^{(2)}) & \cdots & \phi_h(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x^{(i)}) & \phi_2(x^{(i)}) & \cdots & \phi_h(x^{(i)}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x^{(n)}) & \phi_2(x^{(n)}) & \cdots & \phi_h(x^{(n)}) \end{bmatrix}$$

核线性回归的矩阵形式

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

最优参数求解

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \ \rightarrow \boldsymbol{\Phi}^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) = 0$$

$$\rightarrow \boldsymbol{\Phi}^\top\boldsymbol{y} = \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\theta}$$

$$\rightarrow \widehat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}$$

核线性回归的矩阵形式

- 使用线性代数技巧

$$\left(\boldsymbol{P}^{-1}+\boldsymbol{B}^{\top}\boldsymbol{R}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\top}\boldsymbol{R}^{-1}=\boldsymbol{P}\boldsymbol{B}^{\top}(\boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^{\top}+\boldsymbol{R})^{-1}$$

- L2正则化的最优参数

$$\boldsymbol{P}=\frac{1}{\lambda}\boldsymbol{I}_{h\times h}\quad \boldsymbol{R}=\boldsymbol{I}_{n\times n}\quad \boldsymbol{B}=\boldsymbol{\Phi}_{n\times h}$$

$$\begin{aligned}\hat{\boldsymbol{\theta}}&=(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}+\lambda\boldsymbol{I}_h)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y}\\&=\boldsymbol{\Phi}^{\top}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{T}+\lambda\boldsymbol{I}_n)^{-1}\boldsymbol{y}\end{aligned}$$

预测时，不需要使用$\boldsymbol{\Phi}$

$$\begin{aligned}\hat{\boldsymbol{y}}=\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}&=\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}+\lambda\boldsymbol{I}_n)^{-1}\boldsymbol{y}\\&=\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1}\boldsymbol{y}\end{aligned}$$

其中核矩阵$\boldsymbol{K}=\{K(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)})\}$

http://www.ics.uci.edu/~welling/classnotes/papers_class/Kernel-Ridge.pdf

最大似然估计

张伟楠 – 上海交通大学

线性判别模型

判别模型

- 建模预测变量和观测变量之间的关系
- 又名条件模型（Conditional Models）
- 确定性判别模型：$y = f_\theta(x)$
- **概率判别模型**：$p_\theta(y|x)$

带高斯白噪声的线性拟合

$$y = f_\theta(x) + \epsilon = \theta_0 + \sum_{j=1}^{d} \theta_j x_j + \epsilon = \theta^\top x + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$x = (1, x_1, x_2, \dots, x_d)$$

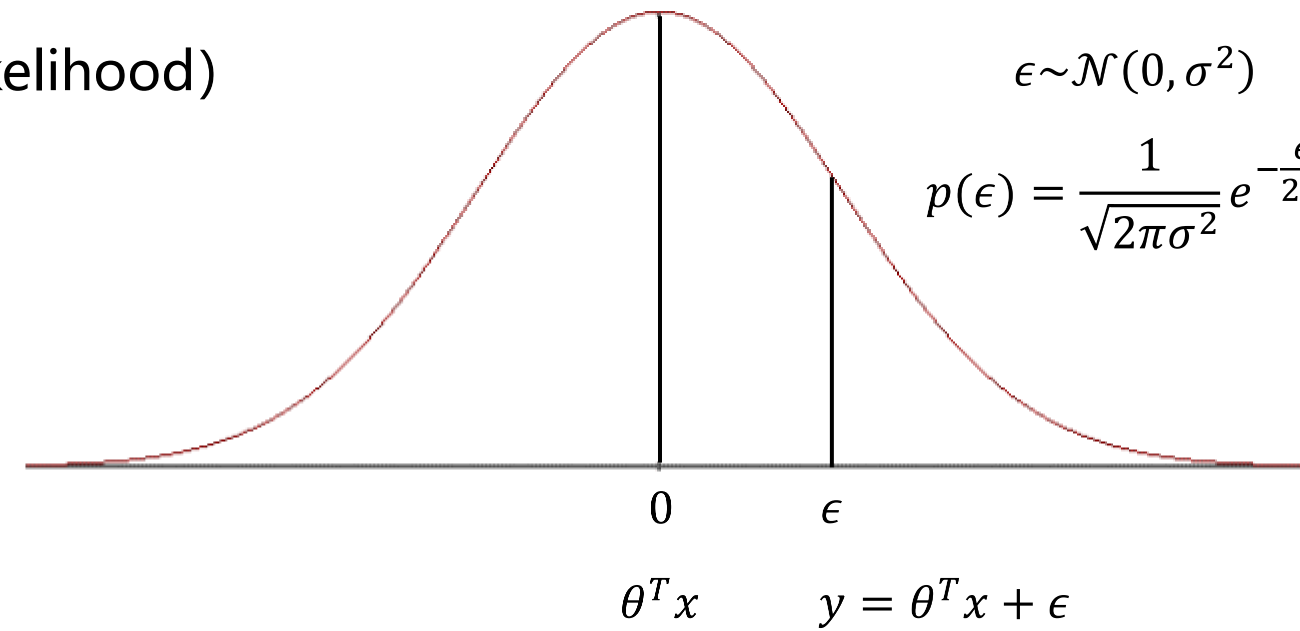
线性判别模型

优化目标

最大似然（likelihood）

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}$$



一个数据点的标签预测似然

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta^{\top}x)^2}{2\sigma^2}}$$

概率判别模型的学习

最大化训练数据的似然

$$\max_{\theta} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}}$$

最大化训练数据的对数似然

$$\log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}} = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}} = -\sum_{i=1}^{N} \frac{(y_i-\theta^\top x_i)^2}{2\sigma^2} + \text{const}$$

$$\min_{\theta} \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2$$

等价于最小均方误差学习

分类指标

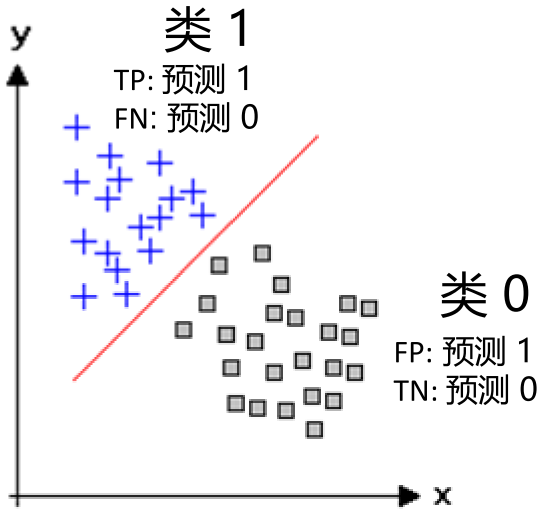
张伟楠 – 上海交通大学

评估指标

预测

		1	0
标签	1	True Positive	False Negative
	0	False Positive	True Negative

- True / False
  - True：预测 = 标签
  - False：预测 $\neq$ 标签
- Positive / Negative
  - Positive：预测$y = 1$
  - Negative：预测$y = 0$



评估指标

预测

标签		1	0
	1	True Positive	False Negative
	0	False Positive	True Negative

精度(Accuracy)

- 分类正确的样本占样本总数的比例

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

评估指标

预测

标签		1	0
	1	True Positive	False Negative
	0	False Positive	True Negative

精确率(Precision)

- 预测为1的样本中标签为1的比例

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

预测

标签		1	0
	1	True Positive	False Negative
	0	False Positive	True Negative

召回率(Recall)

- 标签为1的样本中预测为1的比例

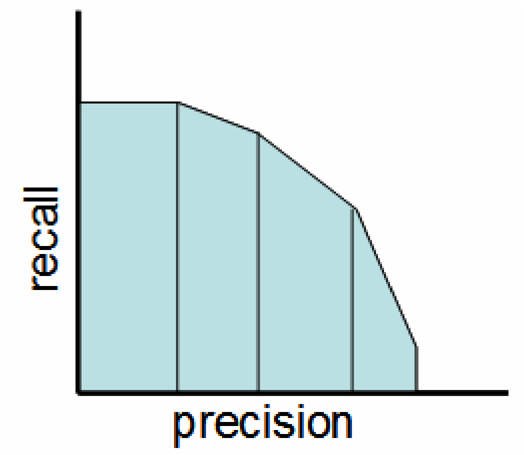
$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

评估指标

精确率和召回率的权衡

$$\hat{y} = \begin{cases} 1, & p_\theta(y=1|x) > h \\ 0, & \text{otherwise} \end{cases}$$

- 阈值越高，精确率越高，召回率越低
  - 极端情况：阈值=0.99
- 阈值越低，精确率越低，召回率越高
  - 极端情况：阈值=0

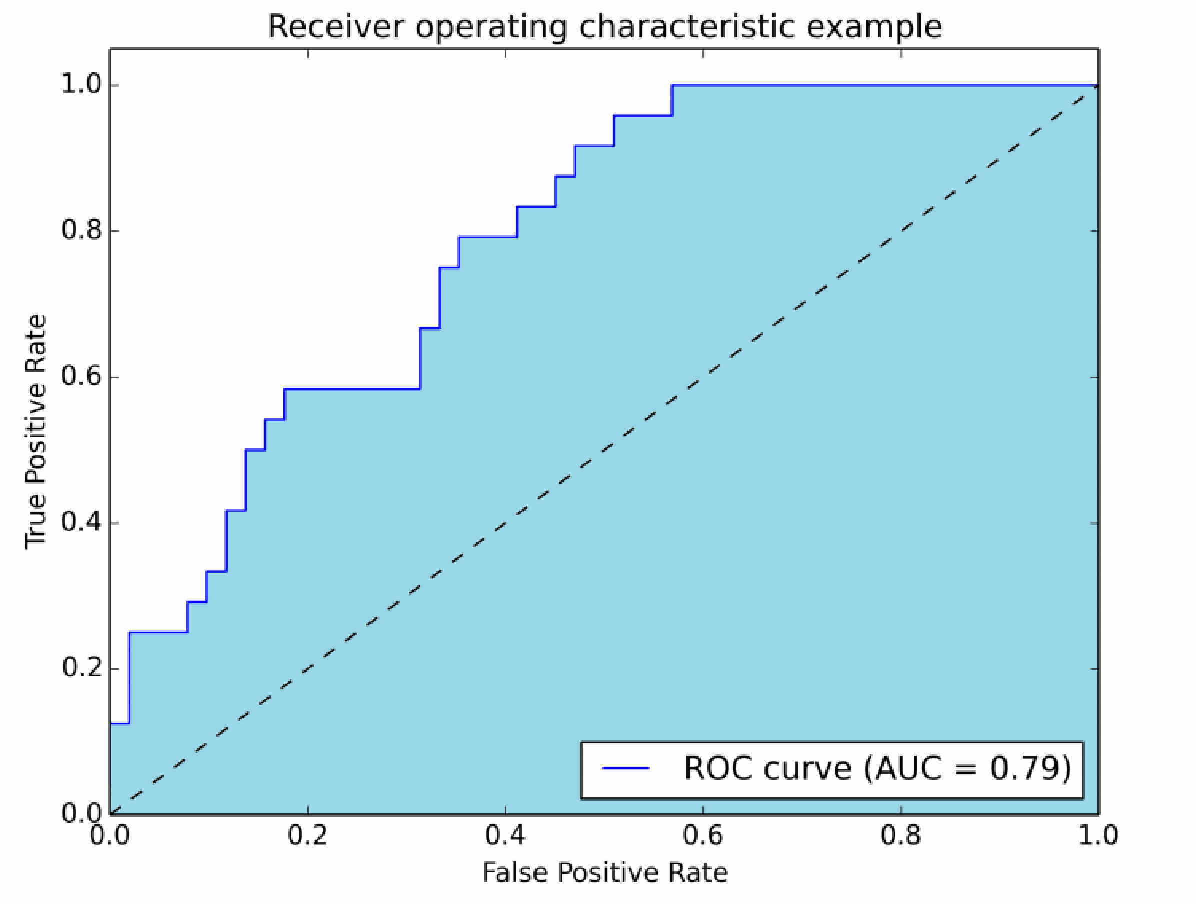


F1分数

$$F1 = \frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}$$

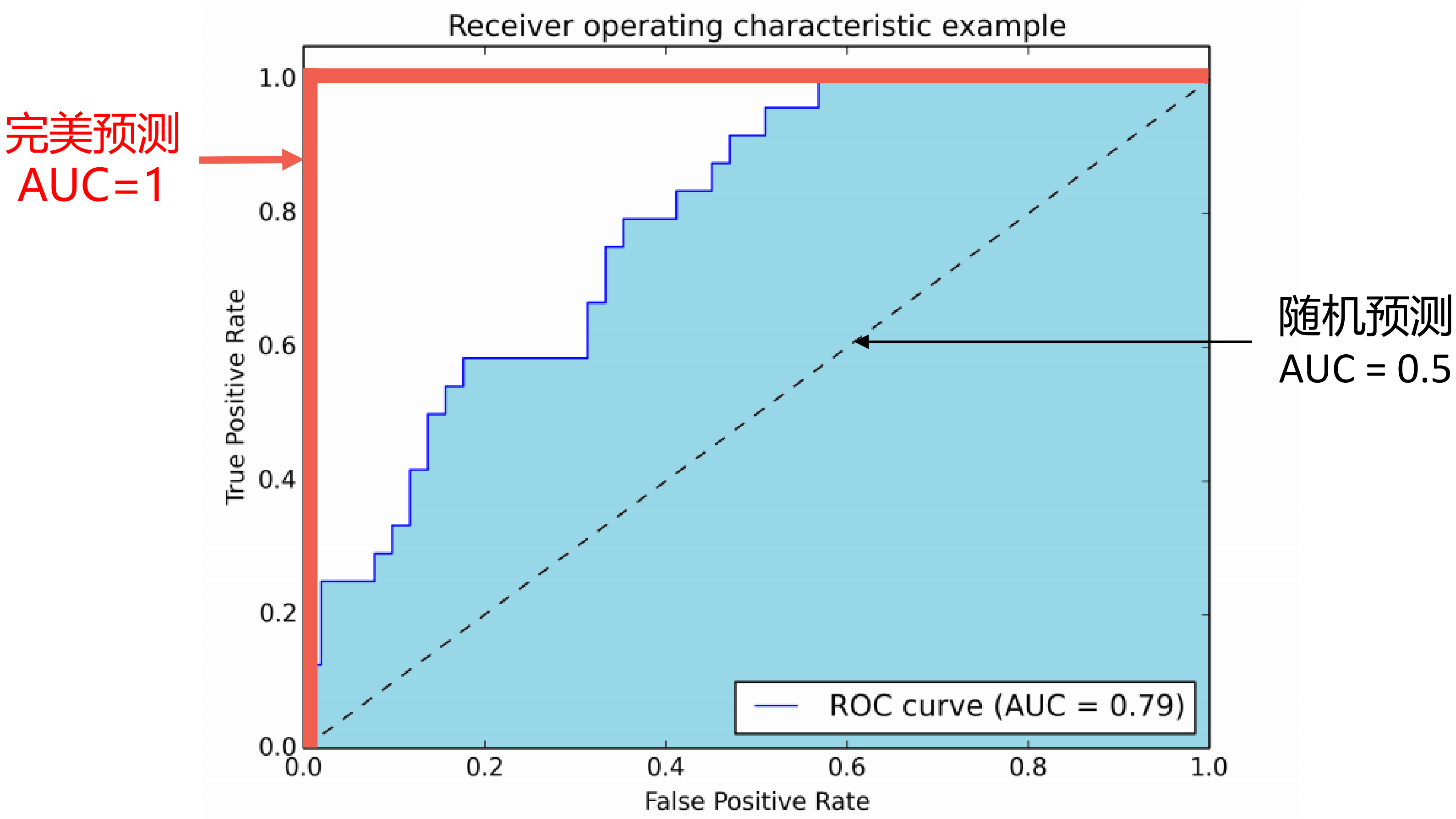
评估指标

- 基于排序的度量：ROC曲线下面积（AUC）



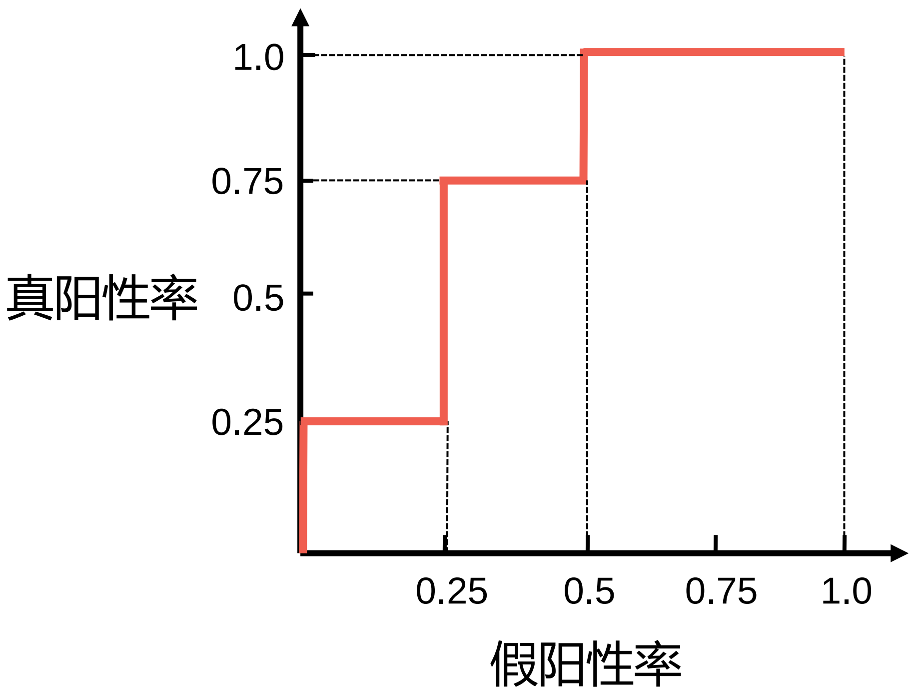
评估指标

- 基于排序的度量：ROC曲线下面积（AUC）



评估指标

AUC计算例子



AUC = 0.75

Prediction	Label
0.91	1
0.85	0
0.77	1
0.72	1
0.61	0
0.48	1
0.42	0
0.33	0

逻辑斯谛回归

张伟楠 – 上海交通大学

目录

Contents

01

二分类

分类问题

给定

- 样本空间$\mathbb{X}$中一个样本x ($x \in \mathbb{X}$)的描述
- 一个固定的类别集：$C = \{c_1, c_2, \cdots, c_m\}$

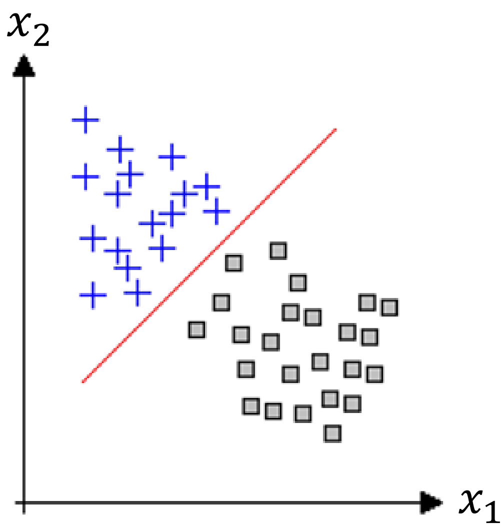
求解

- x的类别:$f(x) \in C$，其中$f(x)$是一个定义域为$\mathbb{X}$，值域为C的类别函数

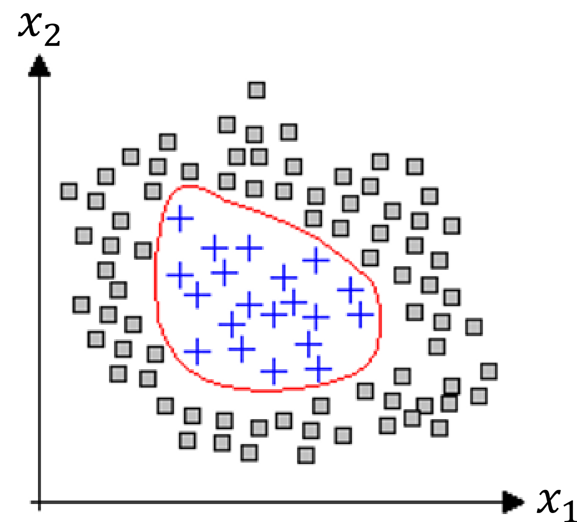
二分类

- 假如类别集是二元的，即$C = \{0,1\}$({错误，正确}，{负，正})，那么这就是二分类问题

二分类



线性可分



线性不可分

线性可分性：是否存在 $ax_1 + bx_2 + c = 0$

使得对于所有的正例：$ax_1 + bx_2 + c > 0$

对于所有的负例：$ax_1 + bx_2 + c < 0$

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型(Conditional Models)
- 分类
  - 确定性判别模型：$y = f_\theta(x)$
    - 对于分类任务不可微分
  - 概率判别模型：$p_\theta(y|x)$
    - 对于分类任务可微分

二分类

$$p_\theta(y = 1|x)$$

$$p_\theta(y = 0|x) = 1 - p_\theta(y = 1|x)$$

损失函数

交叉熵损失

- 离散的情况 $H(p,q) = -\sum_x p(x) \log q(x)$
- 连续的情况 $H(p,q) = -\int_x p(x) \log q(x)\, dx$

分类问题计算交叉熵损失

Ground Truth	0	1	0	0	0
Prediction	0.1	0.6	0.05	0.05	0.2

$$\mathcal{L}(y, x, p_\theta) = -\sum_k \delta(y = c_k) \log p_\theta(y = c_k | x)$$

$$\delta(z) = \begin{cases} 1, & z \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

理解交叉熵

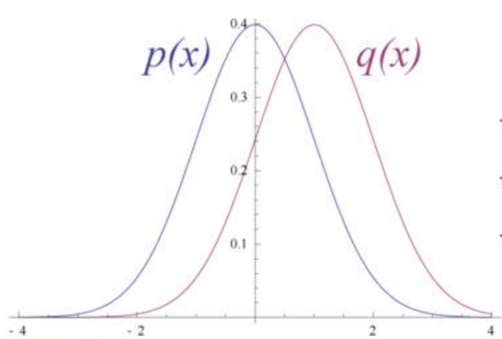
交叉熵损失

- 离散的情况 $H(p,q) = -\sum_x p(x)\log q(x)$
- 连续的情况 $H(p,q) = -\int_x p(x)\log q(x)\,dx$

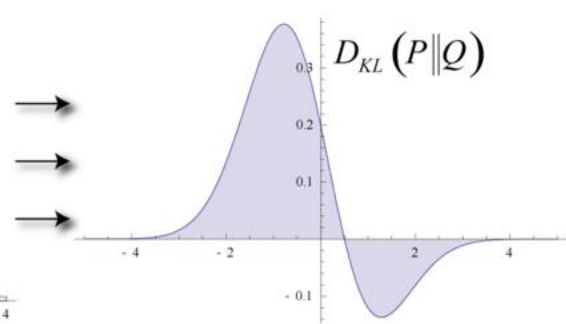
KL散度

$$KL(p,q) = H(p,q) - H(p)$$

- 衡量两个分布的距离：恒大于0，两者相等时为0，不对称
- 离散的情况 $KL(p,q) = \sum_x p(x)\log\frac{p(x)}{q(x)}$
- 连续的情况 $KL(p,q) = \int_x p(x)\log\frac{p(x)}{q(x)}\,dx$



Original Gaussian PDF's



KL Area to be Integrated

二分类的交叉熵

	Class 1	Class 2
真实值	0	1
预测值	0.3	0.7

- 损失函数

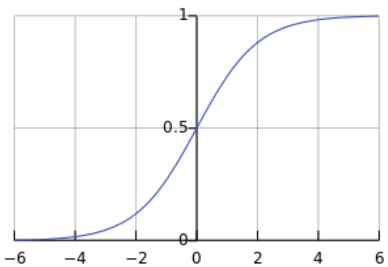
$$\begin{aligned}\mathcal{L}(y, x, p_\theta) &= -\delta(y=1)\log p_\theta(y=1|x) - \delta(y=0)\log p_\theta(y=0|x) \\ &= -y\log p_\theta(y=1|x) - (1-y)\log\big(1-p_\theta(y=1|x)\big)\end{aligned}$$

逻辑斯谛(Logistic)回归

- 逻辑斯谛回归是一个二分类模型

$$p_\theta(y=1|x)=\sigma(\theta^\top x)=\frac{1}{1+e^{-\theta^\top x}}$$

$$p_\theta(y=0|x)=\frac{e^{-\theta^\top x}}{1+e^{-\theta^\top x}}$$



- 交叉熵损失函数

$$\mathcal{L}(y,x,p_\theta)=-y\log\sigma(\theta^\top x)-(1-y)\log(1-\sigma(\theta^\top x))$$

- 梯度

$$\begin{aligned}\frac{\partial\mathcal{L}(y,x,p_\theta)}{\partial\theta}&=-y\frac{1}{\sigma(\theta^\top x)}\sigma(z)\big(1-\sigma(z)\big)x-(1-y)\frac{-1}{1-\sigma(\theta^\top x)}\sigma(z)\big(1-\sigma(z)\big)x\\&=(\sigma(\theta^\top x)-y)x\end{aligned}$$

$$\theta\leftarrow\theta+\eta\big(y-\sigma(\theta^\top x)\big)x$$

$$\boxed{\frac{\partial\sigma(z)}{\partial z}=\sigma(z)(1-\sigma(z))}$$

标签的决定

- 逻辑斯谛回归求出的概率

$$p_\theta(y=1|x)=\delta(\theta^\top x)=\frac{1}{1+e^{-\theta^\top x}}$$

$$p_\theta(y=0|x)=\frac{e^{-\theta^\top x}}{1+e^{-\theta^\top x}}$$

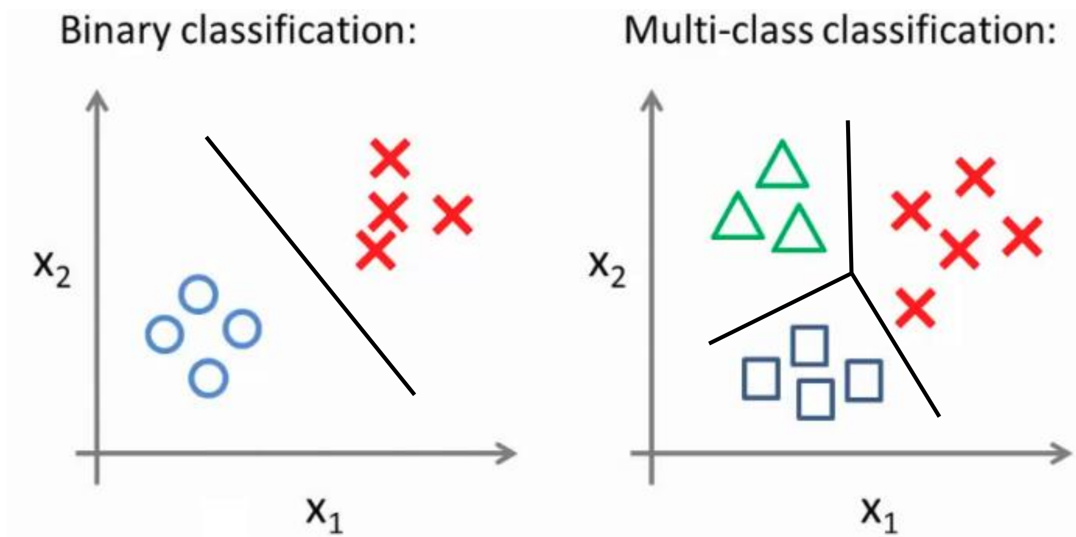
- 设置阈值(threshold)h决定示例最终标签

$$\hat{y}=\begin{cases}1, & p_\theta(y=1|x)>h \\ 0, & \text{otherwise}\end{cases}$$

02

多分类

多分类



多分类交叉熵

$$\mathcal{L}(y, x, p_\theta) = -\sum_k \delta(y = c_k) \log p_\theta(y = c_k | x)$$

$$\delta(z) = \begin{cases} 1, & z \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

真实值	0	1	0
预测值	0.1	0.7	0.2

多类别逻辑斯谛回归

- 类别集

$$C = \{c_1, c_2, \cdots, c_m\}$$

- 预测$p_\theta(y = c_j|x)$的概率

$$p_\theta(y = c_j|x) = \frac{e^{\theta_j^\top x}}{\sum_{k=1}^m e^{\theta_k^\top x}} \quad \text{for } j = 1, \cdots, m$$

- Softmax
  - 参数$\theta = \{\theta_1, \theta_2, \cdots, \theta_m\}$
  - 可以标准化成$m - 1$组参数

多类别逻辑斯谛回归

- 对一个示例的学习（$x, y = c_j$）
  - 最大对数化似然(log-likelihood)

$$\max_{\theta} \log p_{\theta}(y = c_j | x)$$

  - 梯度

$$\frac{\partial \log p_{\theta}(y = c_j | x)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \log \frac{e^{\theta_j^{\top} x}}{\sum_{k=1}^{m} e^{\theta_k^{\top} x}}$$

$$= x - \frac{\partial}{\partial \theta_j} \log \sum_{k=1}^{m} e^{\theta_k^{\top} x}$$

$$= x - \frac{e^{\theta_j^{\top} x}}{\sum_{k=1}^{m} e^{\theta_k^{\top} x}} x = \left(1 - p_{\theta}(y = c_j | x)\right) x$$

逻辑斯谛回归的实践

张伟楠 – 上海交通大学

在线广告中的点击率（CTR）估算

是否点击?

大陆

河南省公安厅彻查"封丘36人入警35人身份不合规"

中封丘县公安局的36名受训人员，35人是公安局内部的文职或临时人员，与"民警必须具备公务员身份"的国家规定不符，引发该局内部

- 上海至成都沿江高铁提上日程 串联长江沿线22城市
- 2016号歼-20原型机曝光 已滑行测试(图)
- 日媒：中国或派万吨海警船巡钓鱼岛 打消耗战
- 外媒：中国开始研制隐身武装直升机 预计2020年交付
- 习近平关于中美关系的十个判断
- 住建部黑臭水沟整治工作指南：9成百姓满意才能达标
- 陕西：职校"校长"让女学生陪酒 学校被撤除
- 揭秘"团团伙伙"的武钢漩涡和落马高管

国际

巴塞罗那200万人游行 呼吁加泰罗尼亚独立(图)

- 李炜光：收税是不公平的恶？
- 许章润：超级大国没有纯粹内政
- 刘昀献：国外政党联系群众的路径研究

时局观

民革中央副主席：中共从未否定国民党抗战作用

- 施芝鸿：文革基础上搞改革致一个时期市场官场乱象
- 朱维群回应争议：尊重民族差异而不强化
- 伊协副会长：穆斯林不应因宗教功修忽视社会责任

领袖圈

奥巴马54岁啦，当7年总统人苍老了头发也

海绵城市 未来之城	谈华山论剑与中国精神
水危机：青岛告急	黑龙江创新驱动三步棋
探访中国绿化博览会	《印记》之江城夜未眠
帝都吸引华人首富	办公环境搜查令
凤凰房产 诚邀加盟	圈层生活尽在凤凰会

精彩视频 凤凰联播台

菲媒曝菲律宾军演针对中国 直指南海生命线

播放数：2602282

用户响应估计问题

问题描述

一个实例数据

Date: 20160320
Hour: 14
Weekday: 7
IP: 119.163.222.*
Region: England
City: London
Country: UK
Ad Exchange: Google
Domain: yahoo.co.uk
URL: http://www.yahoo.co.uk/abc.html
OS: Windows
Browser: Chrome
Ad size: 300*250
Ad ID: a1890
User occupation: Student
User tags: Sports, Electronics

对应标签

Click (1) or not (0)?

预测的点击率CTR (0.15)

One-Hot 二进制（独热）编码

- 标准特征工程范例

x=[Weekday=Friday, Gender=Male, City=Shanghai]

x=[0,0,0,0,1,0,0 0,1 0,0,1,0…0]

稀疏表示: x=[5:1 9:1 12:1]

- 高维稀疏二进制特征向量
  - 通常维度超过1M，甚至1B
  - 极其稀疏

训练/验证/测试 数据

- 范例(LibSVM 格式)

1 5:1 9:1 12:1 45:1 154:1 509:1 4089:1 45314:1 988576:1
0 2:1 7:1 18:1 34:1 176:1 510:1 3879:1 71310:1 818034:1
…

- 训练/验证/测试数据拆分
  - 按时间排序数据
  - 训练：验证：测试 = 8：1：1
  - 随机(shuffle)训练数据顺序

逻辑斯谛回归训练

- 逻辑斯谛回归是二分类模型

$$p_\theta(y=1|x)=\sigma(\theta^\top x)=\frac{1}{1+e^{-\theta^\top x}}$$

- 带有L2正则化的交叉熵损失函数

$$\mathcal{L}(y,x,p_\theta)=-y\log\sigma(\theta^\top x)-(1-y)\log\big(1-\sigma(\theta^\top x)\big)+\frac{\lambda}{2}||\theta||_2^2$$

- 参数学习

$$\theta\leftarrow(1-\lambda\eta)\theta+\eta\big(y-\sigma(\theta^\top x)\big)x$$

  - 仅更新标签中非零项对应的参数

实验结果

数据集

- Criteo TB级数据集
  - 13个数字字段，26个分类字段
  - 2014年24天中取七天连续数据(大约300GB)
  - 79.4M次展示，负采样后点击次数为1.6M
- iPinYou数据集
  - 65个分类字段
  - 2013年的连续十天数据
  - 19.5M次展示，负采样后点击次数为937.7K

实验结果

◘ 性能

Model	Linearity	AUC		Log Loss	
		Criteo	iPinYou	Criteo	iPinYou
Logistic Regression	Linear	71.48%	73.43%	0.1334	5.581e-3
Factorization Machine	Bi-linear	72.20%	75.52%	0.1324	5.504e-3
Deep Neural Networks	Non-linear	75.66%	76.19%	0.1283	5.443e-3

◘ 与非线性模型相比，线性模型具有以下优缺点

- 优点：标准化，易于理解和实施，高效和可扩展
- 缺点：建模局限（特征独立假设），无法探索特征交互

THANK YOU